

# Segmentation Rectification for Video Cutout via One-Class Structured Learning

Junyan Wang, *Member, IEEE*, Sai-Kit Yeung, Jue Wang, *Senior Member, IEEE*, and Kun Zhou, *Fellow, IEEE*

**Abstract**—Recent works on interactive video object cutout mainly focus on designing dynamic foreground-background (FB) classifiers for segmentation propagation. However, the research on optimally removing errors from the FB classification is sparse, and the errors often accumulate rapidly, causing significant errors in the propagated frames. In this work, we take the initial steps to addressing this problem, and we call this new task *segmentation rectification*. Our key observation is that the possibly asymmetrically distributed false positive and false negative errors were handled equally in the conventional methods. We, alternatively, propose to optimally remove these two types of errors. To this effect, we propose a novel bilayer Markov Random Field (MRF) model for this new task. We also adopt the well-established structured learning framework to learn the optimal model from data. Additionally, we propose a novel one-class structured SVM (OSSVM) which greatly speeds up the structured learning process. Our method naturally extends to RGB-D videos as well. Comprehensive experiments on both RGB and RGB-D data demonstrate that our simple and effective method significantly outperforms the segmentation propagation methods adopted in the state-of-the-art video cutout systems, and the results also suggest the potential usefulness of our method in image cutout system.



Fig. 0: Given a keyframe segmentation provided by the user (left), our approach generates accurate object cutout results in subsequent frames fully automatically (middle), which can be used for creating a novel compositing (right).

**Index Terms**—Video cutout, segmentation rectification, one-class structured SVM, object segmentation

## 1 INTRODUCTION

VIDEO cutout, as one of the most successful applications of computer vision for video editing and compositing, has gained much attention from the computer graphics community [1], [2], [3], [4], [5], [6]. While practically useful systems have been developed, some fundamental problems still remain unattended. In this paper, we address video cutout from a newly identified fundamental aspect.

### 1.1 Related works

The latest video object cutout systems [5], [6], [7], [8], [9] comprise three major steps: (1) **Keyframe segmentation**, performing keyframe image segmentation and refinement; (2) **Foreground-background classification**, performing classification on other frames given the keyframe segmentation; (3) **Segmentation refinement**, converting the classifier outputs to final cutout results on non-keyframes. Steps 2 and 3 are usually iteratively applied to subsequent frames until the user creates a new keyframe due to occurrence of visible errors.

The keyframe segmentation step can be done using interactive single image segmentation techniques. In foreground-background classification, with the help of motion estimation, foreground classifiers at different scales are constructed/trained from the segmented object in the current frame, and then applied to other frames. The classifiers generate a soft foreground probability map, which is incorporated into a segmentation model in the segmentation refinement step to generate the final object mask on the non-keyframes.

- Junyan Wang is with Doheny Eye Institute at University of California, Los Angeles, CA 90033, USA  
E-mail: ejywang@ucla.edu
- Sai-Kit Yeung is with the Pillar of Information Systems Technology and Design, Singapore University of Technology and Design, Singapore, 487372.  
E-mail: e-mail:saikit@sutd.edu.sg
- Jue Wang is with Adobe Research, Seattle, WA 98103, USA.  
E-mail: e-mail:juewang@ieee.org
- Kun Zhou is with the State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China, 310058.  
E-mail: kunzhou@acm.org.

Manuscript received ; revised

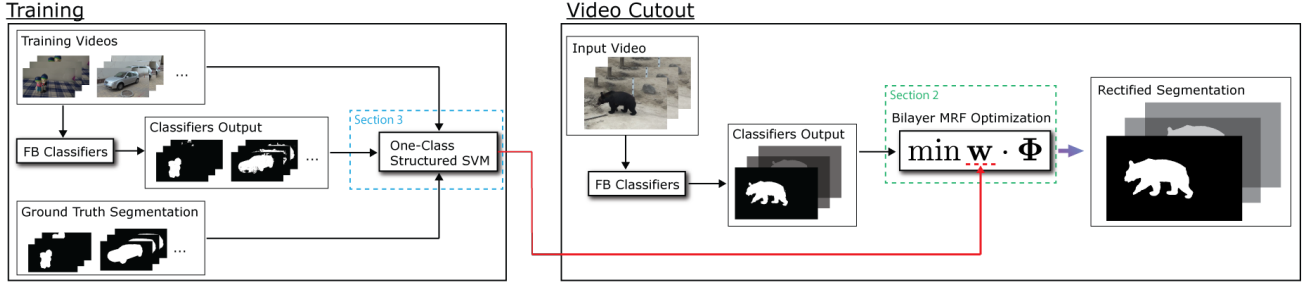


Fig. 2: An overview of our approach.

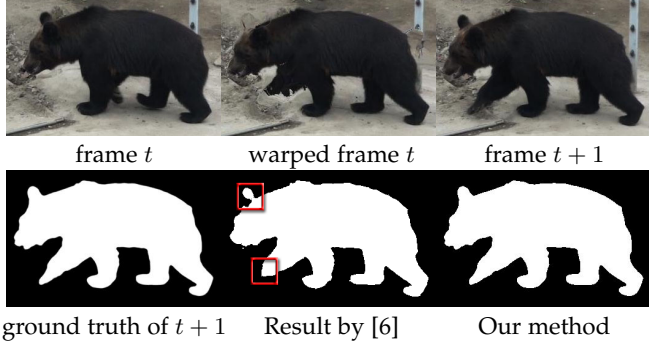


Fig. 1: The need of optimal segmentation rectification.

The Foreground-background classification and segmentation refinement are often considered as a single module called *segmentation propagation*, and most previous works focus only on designing new foreground-background classifiers, while little attention has been paid to the optimal estimation of the segmentation given the classifier output. In Video Snapcut [5], Bai et al. applied the conventional Markov random field (MRF) to the classification output to refine the result. More recently, Zhong et. al [6] applied matting directly after classification. The matting step behaves like random walk segmentation [10] and the latter is closely related to the MRF model [11]. We observe that when the errors from the classifiers tend to bias toward either the false positive or false negative error, the common MRF or matting, which treat the two types of errors equally, would fail to remove the errors satisfactorily. As an example, Figure 1 shows that common method could over shrink the spurious foreground regions produced by the classifier.

More recently, Fan et al. [9] proposed a novel method for propagating segmentation to non-successive frames in order to handle large object displacement. The propagated segmentation mask was refined using geodesic active contour (GAC) model [12] and the level set method [13], rather than MRF with graph cuts. Similar to the MRF based frameworks, GAC with level set method has also not been used to handle the possibly asymmetrically distributed FP and FN. Besides, other video cutout frameworks have been proposed [14], [15], [16]. Our contribution is parallel to these directions towards accurate and user-friendly video cutout.

## 1.2 Contributions

In this paper, we address the possibly asymmetrically distributed FP and FN errors in the output of foreground-background classifier, which we call *segmentation rectifica-*

*tion*. The significance of this subproblem in the context of video cutout is first identified to the best of our knowledge.

Our contribution is twofold. First, we propose a novel bilayer MRF in which the data term can treat the false positive and false negative errors from any given classifier differently using separate weights. Second, we propose a novel one-class structured support vector machine (OSSVM) model to learn the weights, as a computationally more favorable alternative to the conventional two-class structured SVM (2CSSVM) frameworks [17], [18]. We further establish the conditional equivalence between the OSSVM and the conventional (2CSSVM). This theoretical justification of OSSVM is also new in the context of structured learning. Figure 2 illustrates the flowchart of our method.

Our proposed method for segmentation rectification adapts to different classifiers and achieves significant improvement on error reduction over previous methods [5], [6] in segmentation propagation in the experiments. Note that the confidence map adopted in [6] can be used to eliminate unreliable/ambiguous results from classifier output. However, it does not tell where the classifier is overconfident. Our method can remove the error in the classification regardless of the confidence of the classification.

## 1.3 Organization

The rest of the paper is organized as follows. Section 2 derives our bilayer MRF model, Section 3 describes the training process and Section 4 discusses the practical concerns for video cutout. Section 5 details our experiment. Finally, we conclude our work and discuss about the difference between video cutout and other video segmentation tasks in Section 6.

## 2 MODELING SEGMENTATION RECTIFICATION

### 2.1 Segmentation refinement in video cutout

In conventional video cutout systems [5], [6], the classifier output is refined by using the MRF-based segmentation model. The MRF model can be written as:

$$\min_f \sum_{p \in \mathcal{P}} U_p(f_p) + \sum_{\{p, q\} \in \mathcal{N}} V_{pq}(f_p, f_q), \quad (1)$$

where  $p$  refers to a pixel,  $\mathcal{P}$  is the set of all pixels,  $f_p$  is the pixel label and  $\mathcal{N}$  is a neighbourhood system.  $U_p$  and  $V_{pq}$  are the conventional unary and pairwise terms. The unary term can be used to represent the hard constraint given by the user, such as the seeds of foreground and

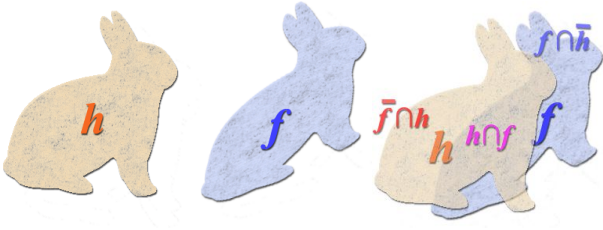


Fig. 3: Illustration of the two types of errors in segmentation propagation.

background regions, or it can be a region model or a shape prior. The pairwise potential is often used to model the object boundaries, and it has also been used to represent advanced priors in segmentation [19].

The unary term for incorporating foreground-background model can be written in the following form:

$$\begin{aligned}
 U_p(f_p) &= \sum_{p \in \mathcal{P}} (f_p - h_p)^2 \\
 &= \sum_{p \in \mathcal{P}} f_p + h_p - 2f_p h_p \\
 &= \sum_{p \in \mathcal{P}} h_p(1 - f_p) + f_p(1 - h_p) \\
 &= \sum_{p \in \mathcal{P}} \bar{f}_p h_p + f_p \bar{h}_p.
 \end{aligned} \tag{2}$$

where  $h_p$  is the classifier output (or probability map that gives the classifier output),  $h_p$  and  $f_p$  are both binary,  $\bar{f}_p = 1 - f_p$  and  $\bar{h}_p = 1 - h_p$ . The unary term is a pixelwise shape distance between the label  $f$  and the classifier output  $h$ . The above equation also provides a decomposition of the unary term, which implies that the unary term above can be naturally related to the two types of errors in segmentation, i.e., the false positives (FP) (background that wrongly considered as foreground)  $\bar{f}_p h_p$ , and the false negatives (FN) (missing foreground)  $f_p \bar{h}_p$ , as illustrated in Figure 3. FPR is defined as

$$FPR = \frac{\text{\# of wrongly classified foreground pixels}}{\text{total \# of pixels}} \tag{3}$$

and FNR is defined as

$$FNR = \frac{\text{\# of wrongly classified background pixels}}{\text{total \# of pixels}}. \tag{4}$$

## 2.2 A case study on the classification error

Here, we conduct a quantitative study on the classification errors produced by the state-of-the-art foreground-background classifier for video cutout [6]. In this study, we perform segmentation propagation using Zhong et al.'s classifier [6] on all consecutive frame pairs in their training dataset, and we compute the false positive ratio (FPR) and false negative ratio (FNR) for each target frame. The quantitative results are shown in Figure 4: the FPR is about 11.22 times greater than the FNR, implying that the two terms in Eq. (2) should not be considered as equally important in the model. This case study disproves the universal fidelity

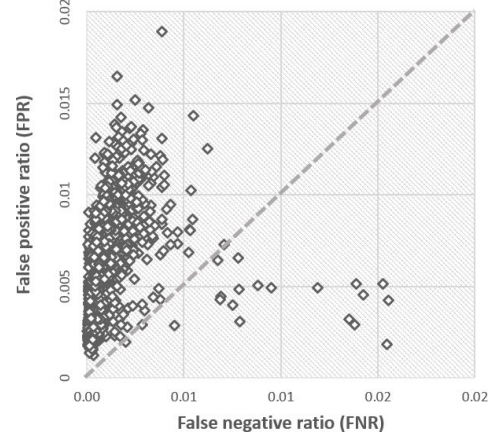


Fig. 4: FPR vs. FNR from Zhong et al.'s FB classifier [6]

of the unary term based on symmetric distance measure in the conventional MRF model. There may be multiple complicated causes of this phenomenon and we omit the in-depth analysis on this specific classifier, since our method is generically applicable to removing errors from any classifier.

## 2.3 A generic shape distance function

Our segmentation rectification method is inspired by the shape-prior based MRF segmentation model [20], [21] in which shape distance is used in the segmentation model for handling occlusion and background clutter. In this work, we view the data term as shape distance and we show why and how we could reformulate this shape distance for segmentation rectification.

From Eq. (2), we observe that the common shape distance used in the data term of the MRF uniquely breaks down into FP and FN with equal weights. To handle the possibly asymmetrically distributed FP and FN errors, we propose the following generic shape distance function to model the relationship between the classifier output and the true segmentation:

$$S_w(f, h) = \sum_{\{p, q\} \in \mathcal{N}_{fh}} w_{pq}^{outside} \bar{f}_p h_q + w_{pq}^{inside} f_p \bar{h}_q, \tag{5}$$

where  $\mathcal{N}_{fh}$  denotes the neighborhood system across  $f$  and  $h$ . In this formulation,  $w_{pq}^{outside}$  and  $w_{pq}^{inside}$  are two unknown data-dependent balancing weights.

The neighborhood system across  $f$  and  $h$  yields a novel graph as visualized in Figure 6. The graph is in a bilayer structure: one layer is defined on the image to represent the unknown segmentation label  $f^t$  at frame  $t$ , and the other layer is used to represent the propagated label  $h^t$ . We have the following observation for different configurations of weights when  $p = q$ :

| weights                              | effect                     |
|--------------------------------------|----------------------------|
| $w_{pq}^{outside} = w_{pq}^{inside}$ | Penalize FP and FN equally |
| $w_{pq}^{outside} > w_{pq}^{inside}$ | Penalize more FP than FN   |
| $w_{pq}^{outside} < w_{pq}^{inside}$ | Penalize more FN than FP   |

This means the FP and FN can be treated differently with proper weights. Besides, by considering all the  $qs$  in  $\mathcal{N}_{fh}$  in





Fig. 5: FP (green box) and FN (red box) from rotobrush on a frame in the “Chinese dance” sequence.

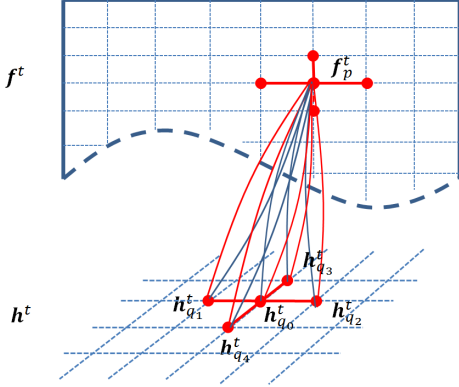


Fig. 6: Graph structure of the proposed bilayer MRF model.  $p$  in  $f^t$  will connect to five neighboring  $qs$  in  $h^t$

the formulation, the noise in the classification result can also be reduced, and the resultant  $S_w$  can be viewed as the local average distance.

## 2.4 The MRF model for segmentation rectification

We can now plug our generic shape decision function in Eq. (5) into the conventional MRF for segmentation propagation to arrive at a novel model for segmentation rectification:

$$\min_{f^t} S_w(f^t, h^t) + \sum_{\{p, p'\} \in \mathcal{N}} \delta_e(f_p^t, f_{p'}^t), \quad (6)$$

where  $h^t, f^t$  are the hypothesis segmentation label and the unknown label at frame  $t$ ,  $\mathcal{N}$  is the neighborhood system for  $f^t$ .

To comply with the standard graph-cuts representation of MRF energy, we may rewrite  $S_w(f^t, h^t)$  as follows

$$S_w(f^t, h^t) = \sum_{\{p, q\} \in \mathcal{N}_{fh}} \delta_s(f_p^t, h_q^t) \quad (7)$$

where  $\mathcal{N}_{fh}$  denotes the neighborhood system for the graph defined on  $h^t$  and  $f^t$ ,  $\delta_s(f_p^t, h_q^t)$  is defined according to the generic shape distance measure, shown in Eq. (5), as:

$$\delta_s(f_p^t, h_q^t) = \begin{cases} w_{pq}^{inside}, & \text{if } h_q^t = 0, f_p^t = 1 \\ w_{pq}^{outside}, & \text{if } h_q^t = 1, f_p^t = 0 \end{cases} \quad (8)$$

where  $w_{pq}^{inside}$  and  $w_{pq}^{outside}$  are to be determined.  $\delta_e(f_p^t, f_{p'}^t)$  in Eq. (6) corresponds to the boundary edge model proposed

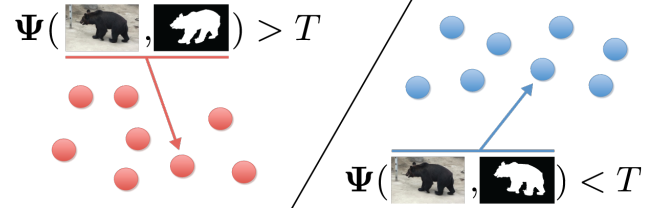


Fig. 7: Idea of SSVM. Red dots represents the “bad” class, and the blue dots represent the “good” class. The black line separating the two classes is the underlying classifier  $\Psi = T$ , where  $T$  is a thresholding constant. The value of  $T$  is implicit in our method.

in [22], which has been shown to be effective for interactive segmentation:

$$\delta_e(f_p^t, f_{p'}^t) = w^{edge} \cdot w_{pq}^e |f_p^t - f_{p'}^t|, \quad (9)$$

where  $w^{edge}$  is the model weight to be determined.  $w_{pp'}^e$  in [22] was defined as:

$$w_{pp'}^e = \begin{cases} \exp(-5I_e(p, p')), & I_e(p, p') \neq 0 \\ 20, & \text{Otherwise} \end{cases} \quad (10)$$

where  $I_e(p, p')$  is 1 if either  $p$  or  $p'$  is on edge. Otherwise, value is 0. We also normalize the weight by  $w_{pp'}^e = w_{pp'}^e / \max_{\{p, p'\} \in \mathcal{N}} (w_{pp'}^e)$ .

The input to the graph cuts solver includes the graph structure, the estimated segmentation of the current image ( $I^t$ ) and the edge map ( $I_e$ ) of the current image. The output is the rectified segmentation ( $f^t$ ).

The energy function in our bilayer MRF model in Eq. (6) can be rewritten in the following compact form:

$$E(f^t | \mathbf{w}, h^t, w_{pp'}^e) = \mathbf{w} \cdot \Psi(h^t, w_{pp'}^e, f^t), \quad (11)$$

where  $\cdot$  is the vector dot product,  $w_{pp'}^e$  is defined in Eq. (10), the weight vector is defined as  $\mathbf{w} = [w^{edge}, w_{pq}^{inside}, w_{pq}^{outside} | \{p, q\} \in \mathcal{N}_{fh}]$ , and

$$\Psi = \begin{pmatrix} \sum_{pp'} w_{pp'}^e |f_p^t - f_{p'}^t|, & \{p, p'\} \in \mathcal{N} \\ \sum_{pq} (1 - h_q^t) f_p^t, & \{p, q\} \in \mathcal{N}_{fh} \\ \sum_{pq} h_q^t (1 - f_p^t), & \{p, q\} \in \mathcal{N}_{fh} \end{pmatrix}, \quad (12)$$

Note that there can be multiple terms for  $\{p, q\} \in \mathcal{N}_{fh}$  as shown in Figure 6. Throughout this paper we mainly consider the following parameterization of  $\mathbf{w}$ :  $\mathbf{w} = [w_1, w_2, \dots, w_{11}]^T = [w^{edge}, w_{p,q_0}^{inside}, w_{p,q_1}^{inside}, w_{p,q_2}^{inside}, w_{p,q_3}^{inside}, w_{p,q_4}^{inside}, w_{p,q_0}^{outside}, w_{p,q_1}^{outside}, w_{p,q_2}^{outside}, w_{p,q_3}^{outside}, w_{p,q_4}^{outside}]^T$ .

## 3 LEARNING THE OPTIMAL MODEL FOR SEGMENTATION RECTIFICATION WITH ONE-CLASS STRUCTURED SVM

Based on the previous observations, it becomes crucial to determine the optimal parameters in the proposed MRF model. One popular method for this task is known as *structured learning* [17], [18], [23], [24]. The basic idea is to consider the MRF parameter learning problem as a classification problem where good segmentations form one class and bad segmentations are the other class. This idea is illustrated in Figure 7. However, this framework requires searching

for negative label samples  $\mathbf{f}_k$ , or the worst case [17], [18], [23], [24], which can be time-consuming. To remove the need for negative samples, we propose to apply the one-class SVM, instead of the conventional two-class formulation [17], to the structured learning problem. The one-class SVM only requires samples from one class, e.g. positive class, for training [25], [26], [27]. Thus, the resultant one-class structured SVM (OSSVM) will also only require the positive samples which are the images paired with ground truth segmentations.

### 3.1 The two-class structured SVM

Before we present our model, we briefly describe the conventional two-class structured SSVM (2CSSVM). We begin with the generic compact form of MRF. The MRF energy for one image can be written as an inner product form w.r.t. the weights  $\mathbf{w}$ :

$$E(f|\mathbf{w}, \mathbf{x}) = \mathbf{w} \cdot \Psi(\mathbf{x}, \mathbf{f}) \quad (13)$$

where  $\mathbf{w}$  is defined in Eq. (11), and  $\mathbf{x} = \{h^t, w_{pp}^e\}$ ,  $\mathbf{f} = f^t$  to be consistent with Eq. (11). Note that this general notations adopted in this subsection allows us to easily extend our approach to other MRF models.

We expect the MRF energy to be as small as possible for the ground truth  $\mathbf{f}^*$  and we expect it to be as large as possible for other non-ideal  $\mathbf{f}$ . This principle can be used for learning the weight vector  $\mathbf{w}$  from data [17], [18], and it can be expressed as:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{N} \sum_{k=1}^N \xi_k \\ \text{s.t.:} \quad & 2\mathbf{w} \cdot \Psi(\mathbf{x}_k, \mathbf{f}_k) - b \geq +1 - \xi_k \\ & 2\mathbf{w} \cdot \Psi(\mathbf{x}_k, \mathbf{f}_k^*) - b \leq -1 + \xi_k \\ & k = 1, 2, 3, \dots, N, \end{aligned} \quad (14)$$

where  $C$  is a constant,  $k$  is the sample id,  $N$  is the number of samples, the constants  $b$  is a bias in conventional decision function and it's unnecessary in segmentation, and  $\xi_k$  is a lack variable that tolerates errors in the training data. In the above model, the margin that separates the positive and negative samples are maximized by minimizing  $\|\mathbf{w}\|^2$ .

Since the positive and negative samples are paired, the respective constraints for each pair of samples can be combined to yield the following constraint:

$$\mathbf{w} \cdot \Psi(\mathbf{x}_k, \mathbf{f}_k) - \mathbf{w} \cdot \Psi(\mathbf{x}_k, \mathbf{f}_k^*) \geq \Delta(\mathbf{f}_k, \mathbf{f}_k^*) - \xi_k \quad (15)$$

Directly combining the constraints in Eq. (14) gives  $\Delta = 1$ . For segmentation problem,  $\Delta$  is a specialized cost and is often set as  $\Delta = \text{mean}(|\mathbf{f}_k - \mathbf{f}_k^*|^2)$  [28]. With the additional requirements of positiveness and boundedness of  $\mathbf{w}$ , the above yields the conventional SSVM of the following form [17], [18], [23], [24].

$$\begin{aligned} \min_{\mathbf{w}, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{N} \sum_{k=1}^N \xi_k \\ \text{s.t.:} \quad & \forall k, \mathbf{w} \cdot (\Psi(\mathbf{x}_k, \mathbf{f}_k) - \Psi(\mathbf{x}_k, \mathbf{f}_k^*)) \geq \Delta_k - \xi_k, \\ & \sum_i w_i = 1, \mathbf{w} \geq 0. \end{aligned} \quad (16)$$

where  $\Delta_k = \Delta(\mathbf{f}_k, \mathbf{f}_k^*)$ . Note that we further imposed a normalization constraint for the weights.

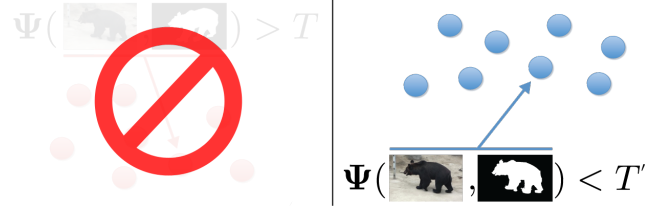


Fig. 8: Idea of OSSVM. The threshold line,  $\Psi = T'$ , is estimated based on the “good” class only. The value of  $T'$  is implicit in our method.

### 3.2 One-class structured SVM

By simply dropping the constraint for non-ideal segmentation  $\mathbf{f}_k$  in Eq. (14) and removing the irrelevant parameter  $b$ , we obtain

$$\begin{aligned} \min_{\mathbf{w}, \varepsilon} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{N} \sum_{k=1}^N \varepsilon_k \\ \text{s.t.:} \quad & \forall k, \mathbf{w} \cdot \Psi(\mathbf{x}_k, \mathbf{f}_k^*) \leq -1 + \varepsilon_k, \\ & \sum_i w_i = 1, \mathbf{w} \geq 0. \end{aligned} \quad (17)$$

where we used  $\varepsilon_k$  instead of  $\xi_k$  to differentiate from the original SSVM formulations. We call this model one-class structured support vector machine (OSSVM), since it does not rely on  $\mathbf{f}_k$ . This optimization problem is well-known as one class support vector machine, and it has been thoroughly studied previously in the classification literature [25]. Here we further establish the rationale of the OSSVM in the context of structured learning.

We observe that the OSSVM, requiring only ground-truth masks, is conditionally equivalent to the conventional two-class SSVM model in which both of non-ideal segmentations and ground-truth segmentations are used. The formal statement is as follows.

**Theorem 3.1.** The OSSVM model in Eq. (17) is identical to the two class SSVM model in Eq. (16) if both of the following conditions are true:

- (I)  $\Delta(\mathbf{f}_k, \mathbf{f}_k^*) = 1$ ;
- (II)  $\forall k, \Psi(\mathbf{x}_k, \mathbf{f}_k) = \mathbf{b}_k$ , where  $\mathbf{b}_k$  is an arbitrary constant vector with equal elements. Its total sum is denoted by  $B_k$ .

There are obviously infinitely many such non-ideal segmentations for many common potentials and the condition allows  $\mathbf{b}_k$  to vary for different  $k$ . We defer its proof to the Appendix.

In a nutshell, this OSSVM model tries to maximize the “margin” from the energy corresponding to all ground-truth samples to the smallest energy produced by the same sample set, such that the margin between the energy of the positive samples and the potential energy of unknown negative samples is also maximized to some extent. We visualize this idea in Figure 8.

**Edge prior.** A price of removing the negative samples is the accuracy of the model. Without negative sample, the training data may not be sufficient to yield satisfactory MRF model. We thus propose to impose a prior during the OSSVM learning. This can be crucial to even two-class SSVM when the ground-truth data itself contains errors. Our

**Algorithm 1: Two-class SSVM learning [24]**


---

**Input** : Training images  $\{\mathbf{x}_k | k = 1, 2, \dots, N\}$  paired with predicted masks  $\{h_k | k = 1, 2, \dots, N\}$  from any classifier and ground truth masks  $\{f_k^* | k = 1, 2, \dots, N\}$

**Output**: Learned weights  $\mathbf{w}^*$  for the MRF potentials

```

1  $\mathbf{w}^0 \leftarrow \vec{1}$ ;
2  $\mathcal{W} \leftarrow \emptyset$ ;
3 forall the Images do
4    $\Psi_k^* \leftarrow \Psi(f_k^* | \mathbf{x}_k, h_k); \backslash \backslash$  By Eq. (12)
5 while Not Converged do
6   forall the Images do
7      $f_k \leftarrow \min_{f'_k} \Delta(f'_k, f_k^*) + \mathbf{w}^j \Psi(f'_k | \mathbf{x}_k, h_k);$ 
8      $\Psi_k \leftarrow \Psi(f_k | \mathbf{x}_k, h_k);$ 
9      $\Delta \Psi_k \leftarrow \Psi_k^j - \Psi_k^*$ ;
10     $\Delta_k \leftarrow \Delta(f_k, f_k^*);$ 
11     $\backslash \backslash$  Cutting plane generation
12     $\mathcal{W} \leftarrow \mathcal{W} \cup \{\Delta \Psi_k, \Delta_k | k = 1, \dots, N\};$ 
13     $\mathbf{w}^{j+1} \leftarrow \begin{cases} \min_{\mathbf{w}, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{|\mathcal{W}|} \sum_{n=1}^{|\mathcal{W}|} \xi_n \\ \text{s.t.: } \forall \{\Delta \Psi_n, \Delta_n\} \in \mathcal{W}, \\ \mathbf{w} \cdot \Delta \Psi_n \geq \Delta_n - \xi_n, \\ \sum_i w_i = 1, \mathbf{w} \geq 0 \end{cases};$ 
14     $\backslash \backslash |\cdot|$  denotes number of elements
15     $j \leftarrow j + 1;$ 
16  $\mathbf{w}^* \leftarrow \mathbf{w}^j;$ 

```

---

prior is that the *edge term is important to segmentation*. Thus, the weight  $w^{edge}$  in Eq. (9), which is actually  $w_1$ , has to be large. This is motivated by the fact that the edge cue is almost always valid, i.e. the final segmentation boundary should always adhere to image edges. To this effect, we propose to maximize the weight on edge features as much as possible. The corresponding one-class SVM model with edge prior for structured learning can be written as follows:

$$\begin{aligned}
& \min_{\mathbf{w}, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{N} \left( \sum_{k=1}^N \varepsilon_k \right) - w^{edge} \\
& \text{s.t.: } \forall k, \mathbf{w} \cdot \Psi(\mathbf{x}_k, \mathbf{f}_k^*) \leq -1 + \varepsilon_k \\
& \sum_i w_i = 1, \mathbf{w} \geq 0,
\end{aligned} \tag{18}$$

### 3.3 Learning algorithms

Both of the SSVM and the OSSVM can be used for learning the weights in our model. We adopt the cutting plane algorithm for two-class SSVM [24]. We include the pseudocode for this method in Algorithm 1 for self-containedness. The pseudocode for our OSSVM learning method is presented in Algorithm 2. After obtaining the weights  $\mathbf{w}^*$ , we can use them in the rectification model shown in Eq. (6) or Eq. (11).

## 4 PRACTICAL CONCERNS IN IMPLEMENTATION

### 4.1 The shrinking bias of graph cut

It is well known that graph cut for MRF model with 2nd-order pairwise potential suffers from the shrinking bias [19], [29]. The recently proposed local foreground-background

**Algorithm 2: OSSVM learning**


---

**Input** : Same as in Algorithm 1

**Output**: Same as in Algorithm 1

```

1 forall the Images do
2    $\Psi_k^* \leftarrow \Psi(f_k^* | \mathbf{x}_k, h_k); \backslash \backslash$  By Eq. (12)
3    $\mathbf{w}^* \leftarrow \begin{cases} \min_{\mathbf{w}, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{N} \left( \sum_{k=1}^N \varepsilon_k \right) - w^{edge} \\ \text{s.t.: } \forall k, \mathbf{w} \cdot \Psi_k^* \leq -1 + \varepsilon_k, \\ \sum_i w_i = 1, \mathbf{w} \geq 0 \end{cases};$ 
4    $\backslash \backslash$  According to Eq. (18)

```

---

classifiers [5], [6] are shown to be able to correct local errors near the object boundary. The shrinking bias falls into this category as it introduces small errors near the boundary. Hence, we propose to feed the results of the rectified segmentation to re-train the foreground-background classifiers on the current frame, and use the updated classifiers to perform classification in the same frame again, to avoid the shrinking bias.

### 4.2 Computational complexity

There exist quite a few efficient algorithms for solving graph cuts, i.e. the max-flow/min-cut problem. The computational time for the Boykov-Kolmogorov (BK) algorithm on a 2 MP image on CPU is about 160 ms, and the GPU implementation of graph cuts can be 2 times faster than on the CPU [30]. The foreground-background classifier we use is the one reported in [6], and its average computational time is about 1.5s for one frame on a PC with quad-core 3.3 GHz CPUs. Optical flows and edge maps can all be precomputed. worth showing due to the page limit. We include them in the supplementary material.

## 5 EXPERIMENTAL RESULTS

In the experiments, we compare our method with state-of-the-art methods for full sequence cutout with the initial keyframe segmentation. We avoid end-to-end system comparison since the interactive segmentation phase, i.e., Step 1, in the video cutout system is out of the focus of this work. We are unable to present all the results worth showing due to the page limit. We include them in the supplementary material.

### 5.1 Experimental Setup

**Datasets.** We mainly evaluate our method on the dataset proposed in Zhong et al. [6]. The main advantage of this dataset is that the ground truth segmentation for each frame has been provided. The training data from Zhong et al.'s dataset contains 15 video sequences in total, 9 for training and 6 for testing. The *Training set* we used in this work contains 2012 frames from their 9 training sequences, and the *Test set* contains 741 frames from their testing sequences.

**Learning the weights.** We use the training set to learn the parameters  $\{w_{pq}^{inside}, w_{pq}^{outside}, w^{edge}\}$  of the bilayer MRF with both the conventional SSVM and our one-class SSVM as presented in section 3.

We considered applying our framework to rectifying the output of two classifiers. One is the Gaussian mixture model

TABLE 1: Learned weights for the bilayer MRF illustrated in Fig. 6. **A** is trained with GMM classifier. **B** is trained with Zhong et al.’s classifier [6].

| <b>w</b> |                  | $w_1$   | $w_2$  | $w_3$  | $w_4$  | $w_5$  | $w_6$  | $w_7$ | $w_8$ | $w_9$ | $w_{10}$ | $w_{11}$ |
|----------|------------------|---------|--------|--------|--------|--------|--------|-------|-------|-------|----------|----------|
| <b>A</b> | 2CSSVM           | 0.52    | 0.01   | 0.01   | 0.01   | 0.01   | 0.01   | 0.048 | 0.089 | 0.098 | 0.098    | 0.093    |
|          | OSSVM w/o prior  | 0.00068 | 0.0059 | 0.0058 | 0.0059 | 0.0059 | 0.0059 | 0.24  | 0.14  | 0.2   | 0.18     | 0.219    |
|          | OSSVM with prior | 0.091   | 0.006  | 0.0059 | 0.006  | 0.006  | 0.006  | 0.25  | 0.12  | 0.17  | 0.14     | 0.19     |
| <b>B</b> | 2CSSVM           | 0.170   | 0.0233 | 0.0424 | 0.0385 | 0.0380 | 0.045  | 0.115 | 0.134 | 0.130 | 0.129    | 0.136    |
|          | OSSVM w/o prior  | 0.0004  | 0.017  | 0.016  | 0.016  | 0.016  | 0.016  | 0.5   | 0.083 | 0.11  | 0.12     | 0.099    |
|          | OSSVM with prior | 0.091   | 0.016  | 0.015  | 0.016  | 0.016  | 0.015  | 0.44  | 0.079 | 0.11  | 0.12     | 0.093    |

where  $\mathbf{w} = [w_1, w_2, \dots, w_{11}]^T = [w_{edge}, w_{p,q_0}^{inside}, w_{p,q_1}^{inside}, w_{p,q_2}^{inside}, w_{p,q_3}^{inside}, w_{p,q_4}^{inside}, w_{p,q_0}^{outside}, w_{p,q_1}^{outside}, w_{p,q_2}^{outside}, w_{p,q_3}^{outside}, w_{p,q_4}^{outside}]^T$

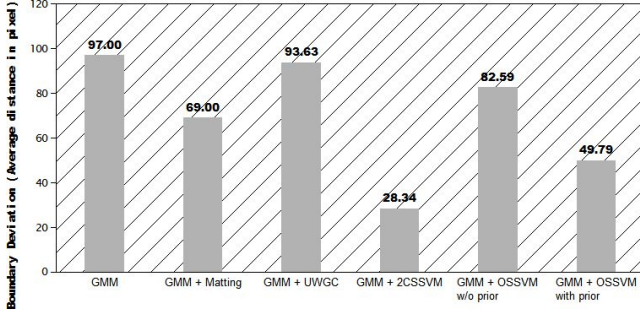


Fig. 9: Effectiveness of our segmentation rectification for GMM classifier.

(GMM), which is a typical global foreground-background (FB) classifier, and the other is the state-of-the-art local FB classifier proposed in [6]. We apply the FB classifiers to all the consecutive frame pairs to generate the classifier output to be rectified. The classifier outputs, the ground truth segmentations, together with the images are then fed into the structured learning framework. Table 1 shows the trained parameters for both GMM and Zhong et al.’s classifier using different learning models. We empirically chose maximum iteration number to be 10 for the training process. The training time for OSSVM is about 0.062 seconds in MATLAB on Intel Core i7-4700MQ Processor, while the training time for the conventional two class SSVM is about 7500 seconds (10 iterations).

It is interesting to note that the weights show *strong asymmetry* structure, and the learned weights would penalize more false positive than false negative. Besides, the weights for GMM are more uniform for both *inside* and *outside* weights. This means the results by GMM is very noisy and strong smoothness is required for rectifying the GMM classifier.

**Evaluation metrics.** We mainly use *boundary deviation* to measure the segmentation performance. It is defined as the average distance from the estimated boundary to the ground truth boundary.

**Methods for evaluation.** We mainly evaluate two variations of our method in the experiments: FB classifier + graph cuts with weights learned by two class structured SVM (2CSSVM), and FB classifier + graph cuts with weights learned by one class structured SVM (OSSVM). We shall call them: **Our method (2CSSVM)**, **Our method (OSSVM)**. We applied our method to GMM based FB classifier and the state-of-the-art FB classifier proposed in [6]. We main compare with the FB classifier + Matting, as adopted in [6], and FB classifier + uniformly weighted graph cuts (UWGC), which is adopted in Rotobrush [5].

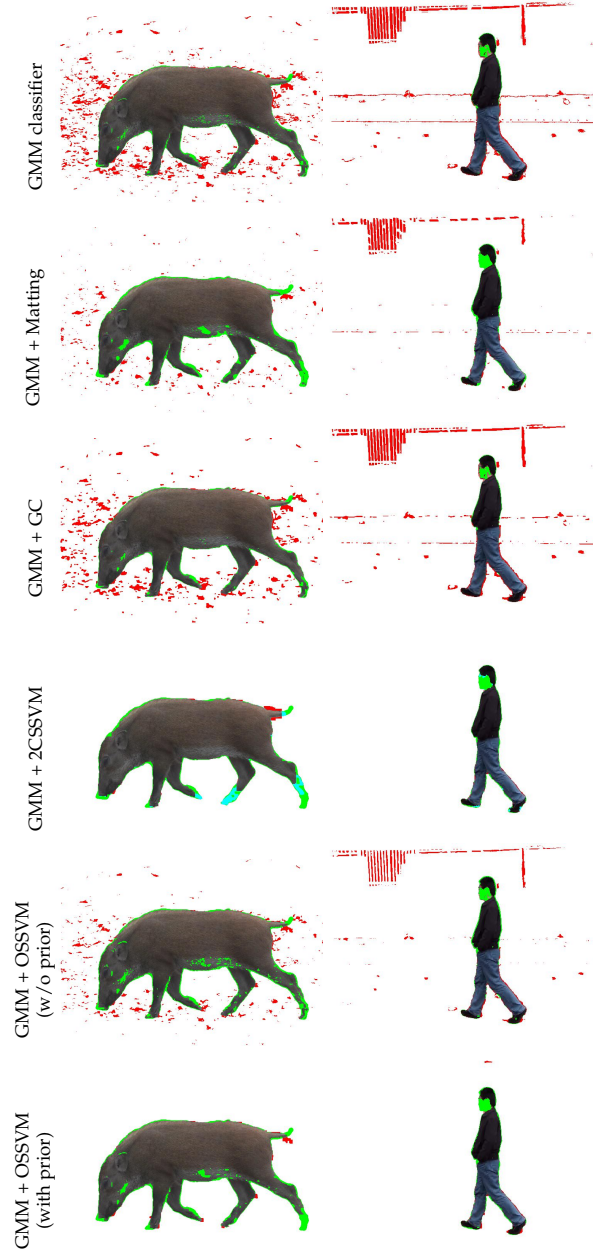
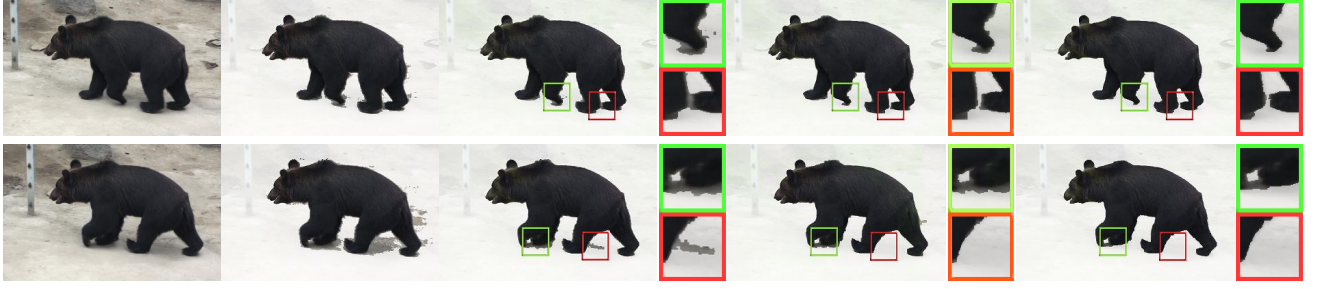
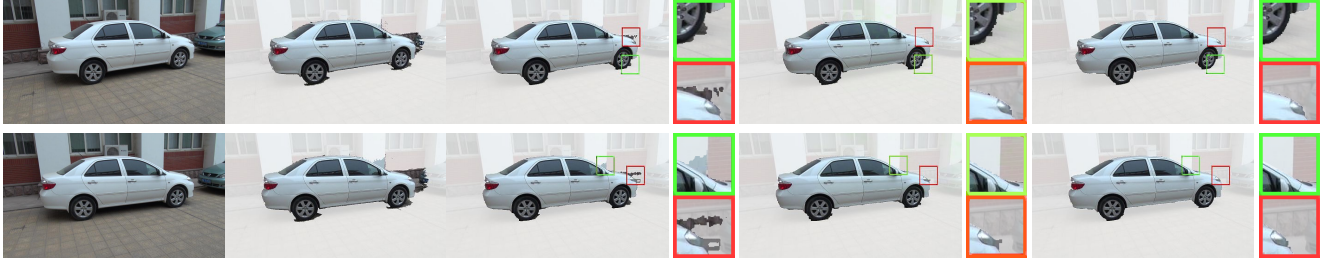


Fig. 10: Comparison of our approach with GMM classifier + other refinement methods. The green regions are the missing foreground regions (FN) and the red regions are the unwanted background regions (FP).





Original image FB Classifier [6] + Matting FB Classifier + UWGC Zoom in Our method (2CSVM) Zoom in Our method (OSSVM) Zoom in  
Fig. 11: Results of fully automatic segmentation propagation for the “Bear” sequence on frames 5 (top) and 9 (bottom), given the same keyframe segmentation. The background is whitened for visualization.



Original image FB Classifier [6] + Matting FB Classifier + UWGC Zoom in Our method (2CSVM) Zoom in Our method (OSSVM) Zoom in  
Fig. 12: Results of fully automatic segmentation propagation for the “Car” sequence on frames 22 (top) and 24 (bottom), given the same keyframe segmentation. The background is whitened for visualization.

## 5.2 Rectifying global classifier

We first apply our framework to the global Gaussian Mixture Model (GMM) classifier for segmentation rectification. Generally, a global classifier like GMM is not suitable to video cutout [5], we conduct the experiment to validate the generality of our framework.

In the segmentation propagation, we train a GMM classifier using the ground truth segmentation on frame  $t$ , and apply it on frame  $t + 1$  as the propagated segmentation, and then apply our rectification approach with different weights, e.g., learned with or without prior as shown in Table 1, for further refinement. The average boundary deviation are summarized in Figure 9. Some typical results are shown in Figure 10. The results suggest that our rectification approach with learned weights can significantly improve the segmentation results generated by the GMM classifier.

Due to the well-known non-local behavior of GMM based classifier, it’s not surprising to see that **2CSVM** and **OSSVM with prior** significantly outperform all other cases. The experimental results shown in this subsection also imply that the errors from GMM for image cutout can be more effectively removed by using our method, compared with the conventional methods based on graph cuts and matting [31], [32], and the 2CSVM outperforms the rest in this case while the OSSVM is the best alternative when computational efficiency in learning is a concern. Note that as shown in the visual results in Figure 10, significant errors can still be observed in the refined output, and this reasserts that global classifier such as GMM alone is not suitable for video cutout.

It is also interesting to note that global classifiers, such as GMM, are commonly adopted in image cutout [31], [32]. Hence, these results suggests that our method may be applied to image cutout as well.

## 5.3 Rectifying local classifier

We further demonstrate the efficacy of our approach for rectifying the state-of-the-art local classifier proposed in [6]. We perform the full sequence propagation on the Test set and Figure 15 shows the quantitative results. Again, we see the segmentation errors by matting and graph cuts with uniform weights accumulates more rapidly than ours. After 10 frames of propagation, matting and graph cut generates 4 times and 2 times more error than our method respectively. We also present the visual results for a “Car” sequence in the Test set in Figure 12, and a “Bear” sequence from the Training set in Figure 11.

We also compare our method with Video SnapCut (Rotobrush) [5] and [6] on additional video sequences. Some of the typical results are shown in Figures 13 and 14. We can observe that the Rotobrush could suffer from the temporal discontinuity while segmentation error could easily accumulate in other methods based on Zhong et al.’s classifier.

## 5.4 Experiment on RGB-D videos

3D movies and videos have now gained increased popularity. An extra depth channel in addition to the RGB channels may be handily available sooner or later. Since our OSSVM framework can be easily extended to handling the extra depth dimension, we also apply our method to RGB-D data for evaluation. The segmentation rectification model for RGB-D data is the same as Eq. (11), except that the potential is defined as

$$\Psi_{RGBD} = \left( \begin{array}{l} \sum_{pp'} w_{pp'}^{eRGB} |f_p^t - f_{p'}^t|, \{p, p'\} \in \mathcal{N} \\ \sum_{pp'} w_{pp'}^{eD} |f_p^t - f_{p'}^t|, \{p, p'\} \in \mathcal{N} \\ \sum_{pq} (1 - h_q^t) f_p^t, \{p, q\} \in \mathcal{N}_{fh} \\ \sum_{pq} h_q^t (1 - f_p^t), \{p, q\} \in \mathcal{N}_{fh} \end{array} \right), \quad (19)$$



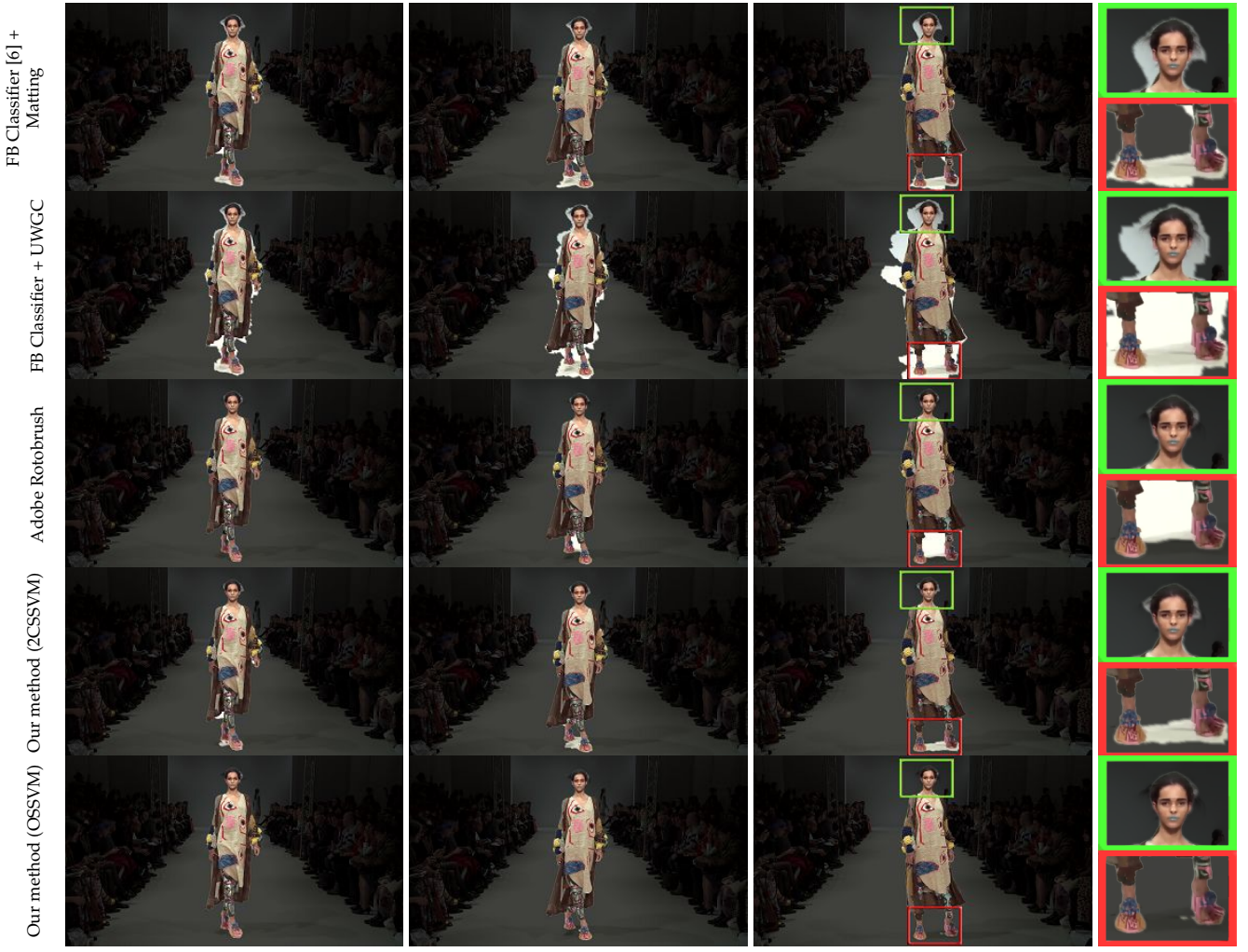


Fig. 13: Results of fully automatic segmentation propagation for the “Catwalk” sequence on frames 1 (left), 6 (middle) and 19 (right), given keyframe segmentation. The background is darkened for visualization.

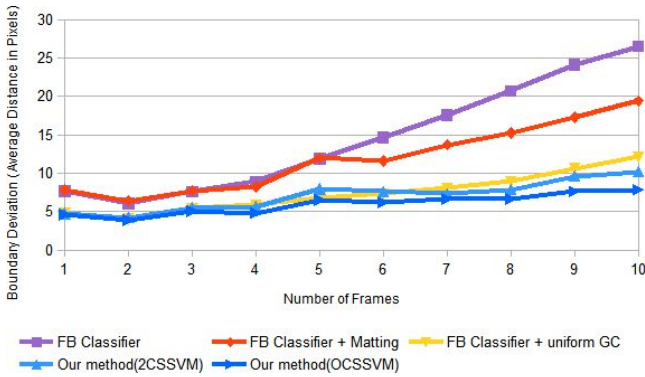


Fig. 15: Error accumulation in segmentation propagation over full sequences on Test set without additional user interaction.

$$w_{pp'}^{eD} = \begin{cases} \exp(-5I_{eD}(p, p')), & I_{eD}(p, p') \neq 0 \\ 20, & \text{Otherwise} \end{cases}, \quad (20)$$

where  $I_{eD}$  is the edge map from depth channel,  $w_{pp'}^{eRGB} = w_{pp'}^e$ , has been defined in Eq. (10).

Similar to the RGB case, the OSSVM model for RGB-D is

$$\begin{aligned} \min_{\mathbf{w}, \varepsilon} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{N} \left( \sum_{k=1}^N \varepsilon_k \right) - w_{RGB}^{edge} \\ \text{s.t.: } \quad & \forall k, \mathbf{w} \cdot \Psi_{RGBD}(\mathbf{x}_k, \mathbf{f}_k^*) \leq -1 + \varepsilon_k, \\ & w_{RGB}^{edge} \leq w_D^{edge}, \\ & \sum_i w_i = 1, \mathbf{w} \geq 0, \end{aligned} \quad (21)$$

in which we added one additional constraint to the model to represent our prior that the edge term from depth is more reliable than that from RGB values and this reformulation does not require additional free parameters.

The dataset we used is from the INRIA 3D movie dataset [33]. Since a number of sequences in the original dataset contain very dark or motion-blurred objects, we select a subset containing 22 sequences with identifiable object boundaries in RGB domain for this experiment. Note that visually identifiable boundaries are required in the context of video cutout and for ground truth delineation. There are a total of 835 frames in the selected subset. Besides, the original dataset only provides the keyframe segmentations. Therefore, we manually cutout each frame.



Fig. 14: Results of automatic segmentation propagation for the “Chinese dance” sequence on frames 32 to 37 (left-right), given the same keyframe segmentation. The background is darkened for visualization.

TABLE 2: Learned weights for RGB-D data

| $\mathbf{w}$           | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | $w_7$ | $w_8$ | $w_9$ | $w_{10}$ | $w_{11}$ | $w_{12}$ |
|------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|----------|----------|
| 2CSVM                  | 0.1   | 0.044 | 0.084 | 0.09  | 0.087 | 0.088 | 0.089 | 0.08  | 0.085 | 0.082    | 0.083    | 0.084    |
| OSSVM with prior = 0.5 | 0.12  | 0.19  | 0.059 | 0.071 | 0.063 | 0.069 | 0.071 | 0.064 | 0.076 | 0.074    | 0.067    | 0.077    |

where  $\mathbf{w} = [w_1, w_2, \dots, w_{12}]^T = [w_{RGB}^{edge}, w_D^{edge}, w_{p,90}^{inside}, w_{p,91}^{inside}, w_{p,92}^{inside}, w_{p,93}^{inside}, w_{p,94}^{inside}, w_{p,90}^{outside}, w_{p,91}^{outside}, w_{p,92}^{outside}, w_{p,93}^{outside}, w_{p,94}^{outside}]^T$

The visual comparison of results from the related methods are shown in Figure 16. From the visual results, we can observe that the results are comparable and our OSSVM on RGB-D outperforms others in general. The quantitative results are shown in Figure 17 and Table 3, which further validated our observation. The first impression is that the overall errors are only around 3 pixels small for many of the methods up to 10 subsequent frames. Besides, we see that our OSSVM on RGB-D and 2CSVM on RGB-D are very comparable and OSSVM is still better than the other methods.

The comparable performance is due to that the object/background motion in the videos in this dataset are often either very small or abrupt, and the method may perform equally good or bad on most of them. The RGB-D

video cutout system would further benefit from a dedicated RGB-D object classifier which is beyond the scope of this paper.

## 5.5 Cross validation for prior weight selection

There is an optional edge prior in our main formulation of OSSVM. The default prior value, which is the coefficient of  $-w^{edge}$ , is 1. From our experiment, we observe that this prior is crucial to the performance. To select the optimal weights, we perform 10-fold cross validation with 7 possible prior weights on a uniform grid  $\{0, 0.5, 1, 1.5, 2, 2.5, 3\}$ , for both of the RGB and RGB-D datasets used in our experiments. The best prior is the one gives overall smallest error in all the 10-fold cross validation.



TABLE 3: Average boundary deviation (ABD) for 10 frame segmentation propagation

| Method     | Zhong et al's Classifier | +Matting in RGB | +Uniform GC in RGBD | +OCSSVM+prior in RGB | +2CSSVM in RGBD | OCSSVM+prior in RGBD |
|------------|--------------------------|-----------------|---------------------|----------------------|-----------------|----------------------|
| ABD(pixel) | 11.01                    | 12.73           | 3.61                | 3.35                 | 3.07            | <b>2.92</b>          |

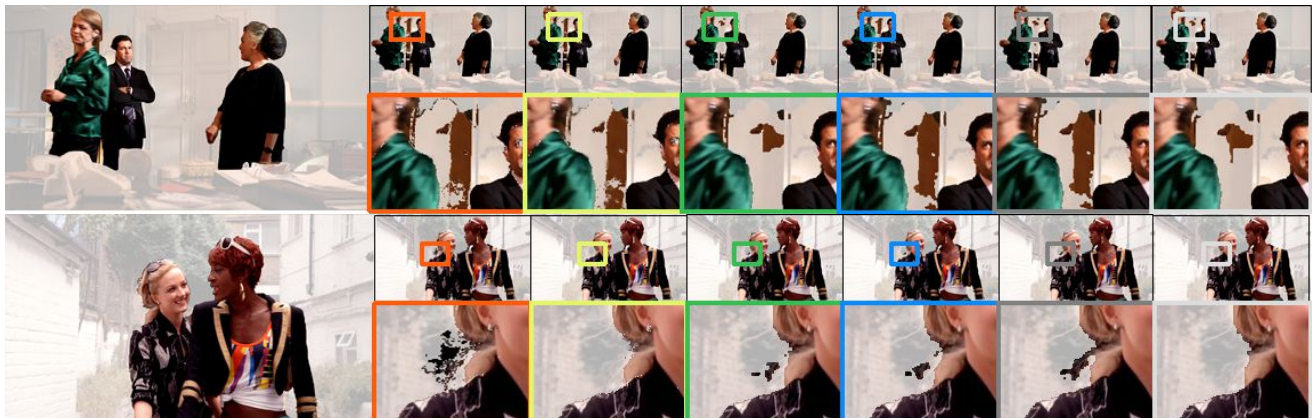


Fig. 16: Segmentation results on two RGB-D sequences (zoom in to see details). In each example, the left most is the initial keyframe segmentation. From the second left to the end are the results by Zhong et al.'s classifier [6], Classifier with Matting, OSSVM with RGB only, uniform weights for GC on RGB-D, 2CSSVM for RGB-D and OSSVM for RGB-D. The top row shows the result on the 5th frame from the keyframe, and the bottom row shows the zoom-in for the boxed regions in the top row.

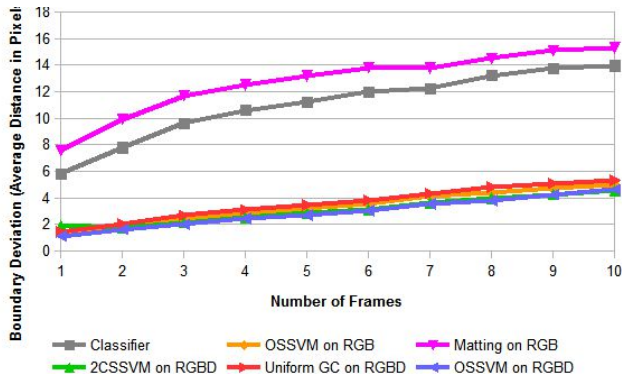


Fig. 17: Error accumulation in segmentation propagation over RGB-D sequences without additional user interaction.

The quantitative results are shown in Figure 18. We can observe that the error in the 10-fold experiments becomes smallest for Zhong et al's dataset if the prior is 1 and the error becomes smallest for the RGB-D dataset if the prior is 0.5. When the prior on the edge weight is too high, the weights on the data term may be too small to produce semantically meaningful results.

## 5.6 Limitations

Although experiments show that our rectification approach can effectively improve the performance of existing video cutout systems, it may fail in challenging cases. A typical failure case is imperfect edge extraction. State-of-the-art edge detection techniques are often reliable, but their results may still contain errors. Such errors may impair the segmentation rectification. See Figure 19 for such an example. The edge map shows that there is a clear strong edge caused by the shadow in-between the legs, causing segmentation rectification to be less effective in this region.

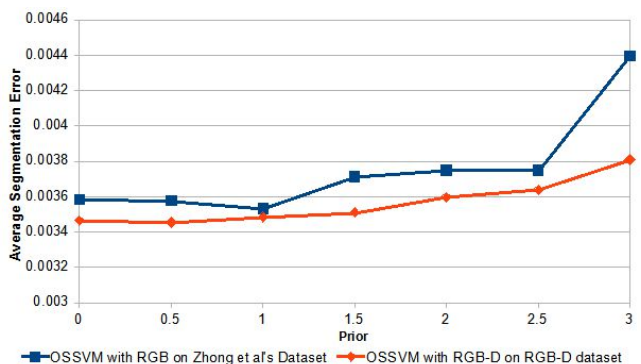


Fig. 18: Cross validation for edge prior weight selection. The average segmentation error is defined as the ratio of wrongly segmented pixels against the total number of pixels.

Another common problem for segmentation propagation is having abrupt changes on the object itself, especially the abrupt emergence of object parts. Figure 20 shows a typical example. Apart from more user interactions, we believe this problem can be solved by using a more sophisticated propagation model, such as a long-term shape prior.

## 6 CONCLUSION AND FUTURE WORK

We propose a novel generic approach to automatically rectify the propagated segmentation in video cutout systems. The core idea of our work is to incorporate a generic shape distance measure in a bilayer MRF framework learned from data to remove the intrinsic bias of the classifier in the segmentation propagation step. This work is motivated by our observation that different classifiers bias toward FP and FN differently, but they were treated equally in the previous video cutout systems. We found that FP and FN can be





Fig. 19: Examples of failure cases I: spurious edges.



Fig. 20: Examples of failure cases II: abrupt emergence of object parts.

treated differently in our bilayer MRF, and the optimal form of the MRF can be learned from the data. Extensive evaluation demonstrates that our approach can significantly improve the state-of-the-art video cutout systems in segmentation accuracy.

There are several vision tasks related to the interactive video cutout problem, such as [34], [35], [36], [37], [38], [39], [40]. In computer vision, those problems can be thought of as shape tracking problem, and errors are generally tolerable. Convenient user interaction is also not a concern to them. In contrast, the interactive video cutout problem does not tolerate visible errors in the segmentation and user-friendly interaction is a crucial concern. It has been proven that the state-of-the-art video cutout frameworks are particularly suitable for the video cutout problem. In most of the works in vision, global optimization frameworks for whole sequences are often adopted. It has been noted in [5] that localized optimization allows user to have better control of the video cutout process. Our method is proposed dedicatedly for video cutout. Its extension for generic video segmentation tasks has yet to be explored.

## APPENDIX

**Proof of Theorem 3.1** We first substitute the two conditions stated in the theorem into the constraint in Eq. (16) and we obtain

$$\begin{aligned} \mathbf{w} \cdot (\Psi(\mathbf{x}_k, \mathbf{f}_k) - \Psi(\mathbf{x}_k, \mathbf{f}_k^*)) &\geq \Delta_k - \xi_k \\ \Leftrightarrow \mathbf{w} \cdot (\mathbf{b}_k - \Psi(\mathbf{x}_k, \mathbf{f}_k^*)) &\geq 1 - \xi_k \\ \Leftrightarrow \mathbf{w} \cdot \Psi(\mathbf{x}_k, \mathbf{f}_k^*) &\leq -1 + \hat{\xi}_k \end{aligned} \quad (22)$$

where we have applied the normalization constraint  $\sum_i w_i = 1$  and we set  $\hat{\xi}_k = B_k + \xi_k$ ,  $B_k = \sum_j [\mathbf{b}_k]_j$ . The above gives us the constraint in the OSSVM formulation in Eq. (17).

The objective function in Eq. (16) can accordingly be rewritten as

$$\begin{aligned} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{N} \sum_{k=1}^N (\hat{\xi}_k - B_k) \\ = \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{N} \sum_{k=1}^N \hat{\xi}_k - \frac{C}{N} \sum_{k=1}^N B_k \end{aligned} \quad (23)$$

where the last term is a constant. By omitting the constant we obtain the objective function in OSSVM Eq. (17).

Lastly, due to the one-to-one correspondence between  $\hat{\xi}_k$  and  $\xi_k$  for any  $k$ , we know that optimizing over  $\mathbf{w}$  and  $\{\xi_k\}$  is equivalent to optimizing over  $\mathbf{w}$  and  $\{\hat{\xi}_k\}$ , which completes the proof. ■

## REFERENCES

- [1] Y.-Y. Chuang, A. Agarwala, B. Curless, D. H. Salesin, and R. Szeliski, "Video matting of complex scenes," in *SIGGRAPH*, 2002.
- [2] A. Agarwala, A. Hertzmann, D. H. Salesin, and S. M. Seitz, "Keyframe-based tracking for rotoscoping and animation," in *SIGGRAPH*, 2004.
- [3] Y. Li, J. Sun, and H.-Y. Shum, "Video object cut and paste," in *SIGGRAPH*, 2005.
- [4] J. Wang, P. Bhat, R. A. Colburn, M. Agrawala, and M. F. Cohen, "Interactive video cutout," in *SIGGRAPH*, 2005.
- [5] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video snapcut: Robust video object cutout using localized classifiers," in *SIGGRAPH*, 2009.
- [6] F. Zhong, X. Qin, Q. Peng, and X. Meng, "Discontinuity-aware video object cutout," in *SIGGRAPH Asia*, 2012.
- [7] X. Bai, J. Wang, and G. Sapiro, "Dynamic color flow: a motion-adaptive color model for object segmentation in video," in *ECCV*, Springer, 2010.
- [8] F. Zhong, X. Qin, and Q. Peng, "Transductive segmentation of live video with non-stationary background," in *CVPR*, 2010.
- [9] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, and B. Chen, "Jumpcut: Non-successive mask transfer and interpolation for video cutout," 2015.
- [10] L. Grady, "Random walks for image segmentation," *IEEE TPAMI*, vol. 28, no. 11, pp. 1768–1783, 2006.
- [11] A. K. Sinop and L. Grady, "A seeded image segmentation framework unifying graph cuts and random walker which yields a new algorithm," in *ICCV*, 2007.
- [12] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contour," *International Journal of Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.
- [13] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations," *Journal of Computational Physics*, vol. 79, pp. 12–49, 1988.
- [14] R.-F. Tong, Y. Zhang, and M. Ding, "Video brush: A novel interface for efficient video cutout," in *CGF*, vol. 30, pp. 2049–2057, Wiley Online Library, 2011.
- [15] L. Zhang, H. Huang, and H. Fu, "EXCOL: an extract-and-complete layering approach to cartoon animation reusing," *IEEE TVCG*, vol. 18, no. 7, pp. 1156–1169, 2012.
- [16] Y. Zhang, Y.-L. Tang, and K.-L. Cheng, "Efficient video cutout by paint selection," *Journal of Computer Science and Technology*, vol. 30, no. 3, pp. 467–477, 2015.
- [17] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin, "Learning structured prediction models: A large margin approach," in *ICML*, 2005.
- [18] I. Tschantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *JMLR*, vol. 6, pp. 1453–1484, Dec. 2005.
- [19] O. Veksler, "Star shape prior for graph-cut image segmentation," in *ECCV*, 2008.
- [20] D. Freedman and T. Zhang, "Interactive graph cut based segmentation with shape priors," in *CVPR*, 2005.
- [21] N. Vu and B. Manjunath, "Shape prior segmentation of multiple objects with graph cuts," in *CVPR*, 2008.
- [22] A. K. Mishra, Y. Aloimonos, L. F. Cheong, and A. Kassim, "Active visual segmentation," *IEEE TPAMI*, vol. 34, pp. 639–653, Apr. 2012.
- [23] M. Szummer, P. Kohli, and D. Hoiem, "Learning crfs using graph cuts," in *ECCV*, 2008.
- [24] T. Joachims, T. Finley, and C.-N. J. Yu, "Cutting-plane training of structural svms," *Machine Learning*, vol. 77, pp. 27–59, Oct. 2009.
- [25] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, pp. 1443–1471, July 2001.
- [26] Y. Chen, X. S. Zhou, and T. S. Huang, "One-class svm for learning in image retrieval," in *ICIP*, 2001.
- [27] L. M. Manevitz and M. Yousef, "One-class svms for document classification," *JMLR*, vol. 2, pp. 139–154, 2002.

- [28] S. Nowozin and C. H. Lampert, "Structured learning and prediction in computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, no. 3–4, pp. 185–365, 2011.
- [29] B. L. Price, B. Morse, and S. Cohen, "Geodesic graph cut for interactive image segmentation," in *CVPR*, 2010.
- [30] V. Vineet and P. Narayanan, "Cuda cuts: Fast graph cuts on the gpu," in *CVPR*, 2008.
- [31] C. Rother, V. Kolmogorov, and A. Blake, "'grabcut': interactive foreground extraction using iterated graph cuts," in *ACM SIG-GRAPH*, 2004.
- [32] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images," in *ICCV*, 2001.
- [33] G. Seguin, K. Alahari, J. Sivic, and I. Laptev, "Pose estimation and segmentation of people in 3d movies," *IEEE TPAMI*, vol. 37, no. 8, pp. 1643–1655, 2015.
- [34] J. Xiao and M. Shah, "Motion layer extraction in the presence of occlusion using graph cuts," *IEEE TPAMI*, vol. 27, no. 10, pp. 1644–1659, 2005.
- [35] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *ICCV*, pp. 1995–2002, IEEE, 2011.
- [36] M. Gong and L. Cheng, "Foreground segmentation of live videos using locally competing lsvm," in *CVPR*, 2011.
- [37] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *ICCV*, pp. 1777–1784, 2013.
- [38] F. Chen, H. Yu, R. Hu, and X. Zeng, "Deep learning shape priors for object segmentation," in *CVPR*, pp. 1870–1877, 2013.
- [39] S. D. Jain and K. Grauman, "Supervoxel-consistent foreground propagation in video," in *ECCV*, pp. 656–671, Springer, 2014.
- [40] A. Khoreva, F. Galasso, M. Hein, and B. Schiele, "Classifier based graph construction for video segmentation," in *CVPR*, pp. 951–960, 2015.